Technical Report
869

The Intelligibility of Natural and LPC-Vocoded Words and Sentences Presented to Native and Non-Native Speakers of English

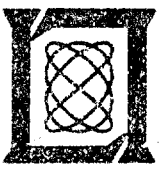
M. Mack
J. Tierney
M.E.T. Boyle

5 July 1990

Lincoln Laboratory

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Lexington, Massachusetts



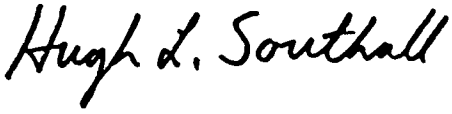
Prepared for the Department of the Air Force
under Contract F19628-90-C-0002.

This report is based on studies performed at Lincoln Laboratory, a center for research operated by Massachusetts Institute of Technology. The work was sponsored by the Department of the Air Force under Contract F19628-90-C-0002.

The ESD Public Affairs Office has reviewed this report, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This technical report has been reviewed and is approved for publication.

FOR THE COMMANDER



Hugh L. Southall, Lt. Col., USAF
Chief, ESD Lincoln Laboratory Project Office

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
LINCOLN LABORATORY

THE INTELLIGIBILITY OF NATURAL AND LPC-VOCODED WORDS AND SENTENCES PRESENTED TO NATIVE AND NON-NATIVE SPEAKERS OF ENGLISH

M. MACK
University of Illinois & Group 24

J. TIERNEY
Group 24

M.E.T. BOYLE
University of California, San Diego

TECHNICAL REPORT 869

5 JULY 1990

LEXINGTON MASSACHUSETTS

ABSTRACT

The experiment reported in the present study was designed to compare the intelligibility of natural and LPC-vocoded linguistic stimuli presented to native and non-native speakers (listeners) of English. Subjects were 20 native speakers of English and 20 native speakers of German who were fluent in English. Three types of stimuli—the Diagnostic Rhyme Test, the Meaningful Sentences Test, and the Semantically Anomalous Sentences Test—were presented in both natural and vocoded conditions.

Results revealed the following: (1) The non-native listeners performed significantly worse than the native listeners in the vocoded condition on the DRT and in the natural and vocoded conditions on the two sentence tests; (2) the effects of listening condition and test type upon response accuracy were nonadditive; (3) the non-native listeners appeared to utilize processing strategies unlike those of the native listeners; (4) the non-native listeners experienced greater recall difficulty than the natives; (5) word frequency affected response accuracy for both subject groups, though somewhat more so for the non-native than for the native listeners; and (6) unlike the native listeners the non-native listeners appeared to exhibit fatigue effects in response to vocoded speech.

These findings provide insight into the role of listening condition and test type in tasks of speech intelligibility, and they reveal differences in the types of response strategies and perceptual learning evinced by native and non-native listeners. In addition, the present study reveals that even moderate amounts of "perceptual loading" can result in serious intelligibility problems for non-native listeners—even when such individuals are quite fluent in the language presented.

ACKNOWLEDGEMENTS

We extend our gratitude to Clifford Weinstein for his continued interest in and support of this work. We also thank Jack Lynch for his critique and suggestions, Linda Nessman and Linda Windhol for their editorial assistance, and Raung-fu Chung for organizational assistance on certain portions of the project.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

1. INTRODUCTION

As the applications for computer-generated speech have increased, so too has the need for assessment of its intelligibility. Within the past two decades, this need has been met by a number of researchers who have attempted to quantify decrements in intelligibility resulting from synthesized speech generated by various algorithms and synthesis procedures (Nye and Gaitenby, 1973, 1974; Voiers, 1977; Pisoni and Hunnicutt, 1980; Pisoni and Koen, 1982; Gold and Tierney, 1983; Luce, Feustel, and Pisoni, 1983; Mack and Gold, 1985; Pisoni, Manous, and Dedina, 1986; Hoover, Reichle, van Tasell, and Cole, 1987).

A consistent finding of such research has been that, although some types of high-quality computer-generated speech may yield quite small decrements in perceptual performance, virtually all computer-generated speech is still less intelligible than natural speech.

It has also long been recognized that the intelligibility of speech is directly related to the test materials and task demands used in its assessment (Miller, Heise, and Lichten, 1951; Kalikow, Stevens, and Elliot, 1977). For example, Pisoni and Koen (1982) found that subjects performed more accurately on the Modified Rhyme Test (MRT) when they were compelled to respond with one of six alternatives (i.e., with a forced-choice closed-set response format) than when they were required to respond freely (i.e., with a free-choice open-set response format). And Pisoni and Hunnicutt (1980) observed that subjects obtained higher scores in response to meaningful than to anomalous sentences. However, in spite of the fact that test materials and task demands may substantially affect intelligibility test results, many researchers continue to administer a single type of test in evaluating coded versus natural speech, thereby possibly underestimating (or overestimating) the extent to which speech intelligibility is affected by coding.

In addition to observing the effects of test materials and task demands, researchers have found a relationship between word frequency and accuracy of word recognition (Savin, 1963; Grosjean, 1980; Mullenix, Pisoni, and Martin, 1989), and between stimulus repetition and recognition accuracy (Miller, Heise, and Lichten, 1951; Pollack, 1959; Clark, Dermody, and Palethorpe, 1985), suggesting that these variables should likewise be considered in tests of speech intelligibility.

Also of importance in intelligibility research is the assessment of the performance of subjects whose native language (L1) is unlike the language of the speech presented. For example, Greene, Pisoni, and Gradman (1985) found that non-native speakers (hereafter listeners) of English performed less accurately than did native listeners on word and sentence intelligibility tests with synthetic speech. Likewise, Mack and Tierney (1987) observed significant differences in the intelligibility performance of English monolinguals and fluent German-dominant German-English bilinguals. In Mack and Tierney's study, subjects were required to provide orthographic transcriptions of natural and vocoded semantically anomalous sentences. Although the stimuli were, acoustically, of high quality (channel-vocoded at 8 kb/s), the bilinguals reproduced only 81.38 percent of the words correctly, while the monolinguals rendered 92.01 percent correctly.

In addition to observing overall differences in the accuracy of native and non-native listeners on tests of speech intelligibility, some researchers have found evidence of differences in *patterns* of response. That is, native and non-native listeners may utilize different sentence processing strategies possibly due to the non-natives' greater difficulty in understanding and/or recalling complex linguistic stimuli (Mack, 1988). For example, Mack and Tierney (1987) found that the German-English bilinguals' error rate in response not only to vocoded but also to *natural* anomalous sentences was very high, suggesting that anomalous sentences may be inherently difficult for non-native listeners. It also provides indirect support for the notion that such listeners must rely heavily on context (e.g., semantic cues) in processing their non-native language in normal communication; for when such cues are absent (as they are in semantically anomalous sentences), non-native listeners perform quite poorly. Mack and Tierney (1987) also found that non-native listeners exhibited proportionately more errors of omission than did native listeners in response to anomalous sentences, suggesting that a test with anomalous sentences places especially strong capacity demands upon the short-term memory processes of non-native listeners.

Other differences in patterns of response in native and non-native listeners were observed by Ozawa and Logan (1989) who found smaller differences between the error rates for natural and processed speech among native than among non-native listeners presented with the Modified Rhyme Test (MRT). They state that their "results suggest that language knowledge and experience may play a more important role in the perception of coded speech than in the perception of unprocessed [natural] speech" (p. 56).

However, such a pattern was not observed by Mack and Tierney (1987) who found that, while vocoded stimuli yielded nearly 7 percent more erroneous words than natural stimuli for native listeners, vocoded stimuli yielded only about 4 percent more erroneous words than natural stimuli for non-native listeners.

Thus, the present study was undertaken in view of previous and sometimes conflicting findings concerning the intelligibility of natural versus computer-generated speech presented to different types of subjects in various listening conditions. Specifically, this study was designed to address the following three questions: (1) To what extent is intelligibility affected when vocoded versus natural stimuli are presented to native and non-native listeners? (2) What is the magnitude of between-test differences when various types of intelligibility tests are utilized with these two groups? (3) Are specific patterns of response similar for native and non-native listeners?

Answers to these questions have important practical applications and interesting theoretical implications. That is, there is an ever-increasing use of computer-generated speech in the military, commercial, and industrial sectors; invariably, some of the recipients of this speech are non-native listeners. The extent to which they can comprehend the speech becomes especially important if rapid and/or accurate comprehension is essential. Hence, data from a variety of intelligibility tests should be examined if generalizations about the efficacy of "real world" communications systems are to be made. And from a theoretical perspective, evaluation of cross-test scores can reveal whether or not there are consistent absolute or relative differences in intelligibility between native and non-native listeners. This, in turn, can reveal whether or not general predictions may be made about the extent to which "non-nativeness" affects intelligibility. In addition, examination of *specific*

patterns of differences between native and non-native listeners may provide insight into non-native processing strategies and the effects of task demand upon performance in different subject groups.

2. EXPERIMENT

2.1 Subjects

Twenty native speakers of English and 20 native speakers of German served as subjects. Subjects were recruited by means of notices placed in the Foreign Languages Building at the University of Illinois at Urbana-Champaign. Additional subjects, recommended by some of the volunteers, were contacted by telephone. Requests were made for native speakers of English and for native speakers of German who were highly fluent in English (and, of course, in German). Subjects were paid $8 for participating in the experiment.

All but two of the native English speakers and two of the native German speakers were undergraduate or graduate students at the University of Illinois. The native speakers of English had been raised in English-language homes and the native speakers of German had been raised in German-language homes.

Demographic and language-acquisition information about the native and non-native listeners was obtained from self-evaluation questionnaires identical to those used in Mack and Tierney (1987; also reported in Mack, 1988). Responses revealed that the native speakers of English had all grown up in various locales in the United States and that 19 of the non-native speakers were citizens of Germany or Austria. (One was a naturalized U.S. citizen.) All of the German-English bilinguals had begun their study of and/or exposure to English between the ages of 10 and 13.

The non-native subjects also completed a self-evaluation language-proficiency questionnaire (with some of the questions taken from a Foreign Service Institute questionnaire) that included questions about their English proficiency (e.g., "Are you afraid that you will misunderstand information given to you in English over the phone?") and a global English proficiency rating ("Rate your overall proficiency in English on a scale of 1 to 10. 1 represents the score of a low beginner and 10 represents the score of a native speaker"). The highest score achievable was 23—the score a native speaker of English would be expected to obtain. (This score included the global proficiency rating.)[1]

The non-native listeners obtained a mean rating of 13.8 (range 10 through 19). Subjects were also asked to report their TOEFL scores (scores on a standardized test of English as a foreign language), but only three reported having taken the test. Of these three, all had obtained a score of at least 600, placing them in approximately the 90th percentile or above.

Within the two different language groups of 20, ten subjects were randomly assigned to one of two listening conditions, natural speech or vocoded speech, yielding ten subjects in each one of four groups—English natives presented with natural speech (ENG-NAT), English natives presented

[1] Although self-assessment is an inherently subjective technique for providing language proficiency data, it is a frequently used approach in bilingual studies and it has been found to correlate significantly with objective language proficiency tests (LeBlanc and Painchaud, 1985).

with vocoded speech (ENG-VOC), German-English bilinguals presented with natural speech (GE-NAT), and German-English bilinguals presented with vocoded speech (GE-VOC). Examination of the global self-rating scores of the two non-native groups revealed that the GE-NAT subjects had a mean rating of 7.4 (range 5 through 9), while the GE-VOC subjects had a mean rating of 7.1 (range 6 through 8).

The mean ages of the subjects in each group were as follows: ENG-NAT, 26 years; ENG-VOC, 28 years; GE-NAT, 28 years; GE-VOC, 29 years. There were four males and six females in each group except the ENG-VOC group in which there were six males and four females.

2.2 Stimuli

Three intelligibility tests were used in the present experiment. These were the Diagnostic Rhyme Test (DRT), the Meaningful Sentences Test (MST), and the Semantically Anomalous Sentences Test (SAST).

The DRT is a test in which subjects must identify which member of a minimal pair (e.g., "pond"–"bond") has been presented. The DRT was selected for the present study because it has been widely used for assessing the intelligibility of computer-processed and synthesized speech and has been established as a NATO and Department of Defense test standard for evaluating narrowband speech processors. It is also a comparatively simple test, placing minimal demands on processing capacity and short-term memory.

The MST, a phonemically balanced test of meaningful sentences devised for the present study, was used because it requires subjects to utilize some of the same mechanisms used in normal sentence processing. Thus it satisfied the need for a "reasonably natural" intelligibility test (Kalikow, Stevens, and Elliot, 1977). Furthermore, it was believed that the open-set free-response format of the MST would place greater demands upon processing and recall mechanisms than the DRT and hence could reveal significant between-group differences which might not be apparent in DRT scores.

The SAST, a phonemically balanced test of anomalous sentences was identical to that used in two previous studies (Mack and Gold, 1985; Mack and Tierney, 1987). Because the SAST sentences are almost completely lacking in semantic or contextual cues, it was believed that they would provide the strongest test of between-group differences.[2]

The DRT used in the present study contained 232 words (appendix A). Thus, 9280 words were scored (232 words × 40 subjects). There were 20 English phonemes in word-initial position, each occurring from 2 to 26 times. Stimuli were real words, with the exception of "daw," "vox," "foo," "bon," and—possibly for American English speakers—"zed."

Each sentence test contained 57 sentences and a total of 383 words (appendices B and C). Thus, 30,640 words were scored (383 words × 2 tests × 40 subjects). The meaningful and anomalous

[2] Words in the sentence tests were phonemically balanced to permit analysis of phonological errors. This analysis is not reported in the present study.

sentences were constructed using nouns, adjectives, and verbs in six- or seven-word grammatical sentences. SAST sentences were devised by randomly selecting and then pseudorandomly ordering a set of previously selected nouns, verbs, and adjectives. MST sentences were devised by selecting words not used in the SAST (which was devised prior to the MST) and which could be ordered to produce sentences which two of the experimenters (MM and MB) agreed were meaningful.

In each sentence test, word-initial consonants were phonemically balanced, with each one of 19 English consonant phonemes occurring 15 times—6 times in nouns, 6 times in adjectives, and 3 times in verbs. No consonant clusters occurred in word-initial position. Sentences were of the form S → NP + VP where NP → (art +) adj + noun, and VP → verb + art + adj + noun. (An article was not included in 16 of the sentences in each test to reduce some of the syntactic predictability of the sentences without fundamentally altering their structure.) All words were mono- or bisyllabic or were pronounced as such in the dialect of the speaker. No words were repeated within a test or across tests.

It must be noted that the phonological and syntactic constraints imposed upon the selection of words and sentences in the MST resulted in the use of some low-frequency words and some fairly unusual sentences. So to confirm that the sentence tests did indeed differ with respect to their meaningfulness, the present researchers obtained sentence ratings for the MST and SAST from 25 native speakers of English. Raters were undergraduate and graduate students at the University of Illinois who were presented with all MST and SAST sentences in random order, printed on response sheets. (Three different randomization orders were used.) Raters were told to place an X next to any "nonsense" sentence, and they were given five practice sentences with feedback immediately prior to the experiment.

Results revealed large differences in the ratings given to the MST and SAST sentences. That is, an average of only 2.84 (4.98 percent) of the meaningful sentences were rated as anomalous, and 10 (17.54 percent) of the anomalous sentences were rated as meaningful.

In addition, only two of the meaningful sentences were rated as anomalous by 33 percent or more of the judges, and only three of the anomalous sentences were rated as meaningful by 33 percent or more of the judges. These findings indicated that there were—at least for native speakers of English—clear differences between the MST and SAST with respect to the meaningfulness of nearly all sentences.

Word-frequency counts of all words were obtained to determine the comparability of words in the three tests and to provide information about frequency that would later be used in the analysis of intelligibility test results. Word-frequency counts were obtained from the American Heritage Intermediate Corpus—a corpus of over 86,000 word types obtained from a sample of over five million word tokens. Frequency counts of the words in the present study were based upon a reported value of U, where U is equal to the estimated frequency-per-million tokens derived from overall frequency with an adjustment for dispersion over subject categories. (U is believed to be a more accurate reflection of word frequency than simple frequency—i.e., it reflects the frequency-per-million in a corpus of indefinitely large size.)

Analysis of the DRT words revealed a mean word frequency of 155.59 with a range of 0.01 to 3630.80. The percentage of words in the DRT occurring with a frequency of 20 or more was 37.04. Content words in the MST had a mean frequency of 56.55, with a range of 0.01 to 1344.10. The percentage of words in the MST occurring with a frequency of 20 or more was 35.14. Content words in the SAST had a mean frequency of 66.32, with a range of 0.02 to 1228.50. The percentage of words in the SAST with a frequency of 20 or more was 41.37. Thus, in spite of the large frequency ranges across tests, 35 to 41 percent of the words in all three tests occurred at least 20 times per million words.[3]

For later analysis of intelligibility performance as a function of word frequency, each word in the DRT, MST, and SAST was also denoted as belonging to one of three categories—low, mid, or high frequency. Categories were obtained by ordering all words in each test according to frequency, then denoting the bottom third as low frequency, the middle third as mid frequency, and the top third as high frequency.

All stimuli were tape recorded by one of the experimenters (MM) in a sound-attenuated room with high-quality tape-recording equipment. Words and sentences were recorded as they were read at a normal speaking rate and with normal intonation and amplitude. Production of all stimuli was error free. In the DRT, the onset of the words occurred at 3-sec intervals, with an interblock interval of 10 sec after every 29th word. In the MST and SAST, the onset of the sentences occurred at 20-sec intervals with no interblock intervals.

For generating synthetic speech, a master tape recording of the words and sentences was used as input to a 2.4-kb/s Linear Predictive Coding (LPC) vocoder at the MIT Lincoln Laboratory. The LPC vocoder was selected because it is the DoD standard speech coder at a transmission rate of 2.4 kb/s (Tremain, 1982). The speech was analyzed and resynthesized as in normal transmission through the LPC vocoder. The speech was internally sampled at 8 kHz and had an approximate bandwidth of 3.8 kHz. It was then recorded on reel-to-reel magnetic tape for later presentation.

2.3 Procedure

Subjects were tested in groups in the Language Laboratory at the University of Illinois. Test stimuli were presented on a Tandberg TB5200 tape recorder whose output was directed to individual listening consoles. Subjects heard the stimuli over Tandberg stereo headphones with the amplitude set at a comfortable listening level. (They could also change the amplitude at their own listening consoles at any time if they wished.)

[3] As is apparent, there were two major differences between the DRT and the sentences tests. That is, the DRT was not phonemically balanced and the DRT words were all monosyllabic. They were also, on the average, more common than the sentence-test words. While it would have been desirable to have maintained greater comparability between the DRT and the sentences tests, it was recognized that there were advantages to using the tests as designed because the DRT is widely used and well understood, and because the SAST had been administered previously.

Subjects were given booklets in which to provide their answers. For the DRT, a list of minimal pairs (e.g., "dill-gill") was printed, and subjects were told to draw a line through the member of the pair that they heard. For the sentence tests, lines numbered 1 through 57 were printed, and subjects were told to write each sentence as accurately as possible as soon as it was presented. For all tests, subjects were told to guess if they were uncertain and they were encouraged *not* to leave blank items. Tests were presented in the order DRT, MST, and SAST. A break of ten minutes was provided between each test. The total session lasted approximately 90 minutes.

2.4 Data Analysis

2.4.1 Overall Error Analysis

Subjects' DRT data were scored by one of the experimenters (MM), and orthographic transcriptions of the sentences were scored by another (MB). All MST and SAST response sheets were subsequently checked by MM who found a high level of agreement (over 98 percent) between her judgments and those of MB. Where there were differences in the judgments of the two scorers, MM determined how an item was to be scored.

The DRT error analysis was quite straightforward and consisted of counting as one error any incorrect item. (No subjects left any items blank.) In addition to the simple (uncorrected) DRT error scores, DRT scores with correction for guessing were derived, as in Voiers, 1977[4]:

$$S = \frac{100(R - W)}{T}$$

where S = "true" percent correct, R = number of correct items, W = number of incorrect items, and T = total number of items. In the sentence-test error analysis, a word was considered incorrect if it was an inaccurate transcription of the stimulus word. Care was taken *not* to count a misspelling as an error if the pronunciation of the word it spelled was identical to that of the stimulus item. Thus, spellings such as "ladder" for "latter" or "tailer" for "tailor" were counted as correct. The substitution, omission, or insertion of a phoneme or bound morpheme rendered a response erroneous (e.g., "nail" → "male,"; "chili" → "chill"; "hurt" → "hurts"), as did the substitution or omission of a content or function word (e.g., "Zelda" → "failure"; "a" → "the"; "thumped" → 0). An erroneous response in which the syllabic structure of the stimulus word was retained was counted as a single error even if two words resulted (e.g., "raging" → "rage and"). Because the number of lexical insertions was negligible, these were not included in the error analysis.

[4] DRT scores are customarily corrected for guessing. However, Gronlund (1985) recommends such correction when individuals do not have sufficient time to complete a test and when they have been instructed that there will be a penalty for guessing. Because these conditions are not applicable to subjects in the present study, corrected *and* uncorrected scores have been reported, and tabular and graphic DRT data represent uncorrected scores.

2.4.2 Analysis of Other Types of Errors

In addition to the tabulation of erroneous words, other analyses were conducted. These analyses included: (1) number of errors associated with each within-sentence position; (2) number of omitted words; (3) number or errors associated with words of low, mid, and high frequency; and (4) number of erroneous words produced in the first and last 15 test sentences. It was believed that these analyses might reveal differences in the natives' and non-natives' processing strategies and/or recall abilities, their sensitivity to the frequency of lexical items in English, and their ability to learn from repeated presentations of natural and vocoded stimuli.

2.5 Results

2.5.1 Overall Errors

2.5.1.1 Erroneous Words

An identical pattern of response emerged for the subject groups across all three intelligibility tests. That is, in terms of numbers of overall errors on the DRT, MST, and SAST, the four subject groups fell in the following rank order, from lowest to highest number of errors: ENG-NAT, GE-NAT, ENG-VOC, and GE-VOC.

Diagnostic Rhyme Test results revealed that all subjects had fairly high scores in terms of percent correct, with all groups scoring above 90 percent [Figure 2-1(a)]. The ENG-NAT group had a mean score of 99.14 percent, while the GE-NAT group had a mean of 98.19 percent. The ENG-VOC group had a mean score of 94.70 percent, and the GE-VOC group had a mean of 92.37 percent. With correction for guessing, scores were as follows: ENG-NAT = 98.28 percent, GE-NAT = 96.38 percent, ENG-VOC = 89.40 percent, and GE-VOC = 84.74 percent. The raw error data suggests somewhat larger differences among the subject groups than do the percentages (Table 2-1). That is, the ENG-NAT group had a mean of 2.0 errors, while the GE-NAT group had a mean of 4.2. The ENG-VOC group had a mean of 12.3 errors, and the GE-VOC group had a mean of 17.7.

A two-way repeated-measures ANOVA with one-between and one-within subjects factor was used in the statistical analysis of the DRT data. It revealed a significant main effect for language group [$F(1, 36)$, $p < 0.005$] (i.e., the non-natives had significantly more errors than the English natives), a significant main effect for listening condition [$F(1, 36)$, $p < 0.0001$] (i.e., there were significantly more errors in the vocoded than in the natural condition), and no significant language-group $\times$ listening-condition interaction. A Tukey post hoc test revealed no significant difference between the DRT scores of the ENG-NAT and GE-NAT groups. There were, however, significant differences ($p < 0.01$) in the mean DRT scores for the ENG-VOC and the GE-VOC groups, the ENG-NAT and ENG-VOC groups, and the GE-NAT and GE-VOC groups. [5]

[5] Due to the difference in total number of items on the DRT versus the two sentence tests, separate parametric statistics were carried out on the DRT data and on the sentence-test data.

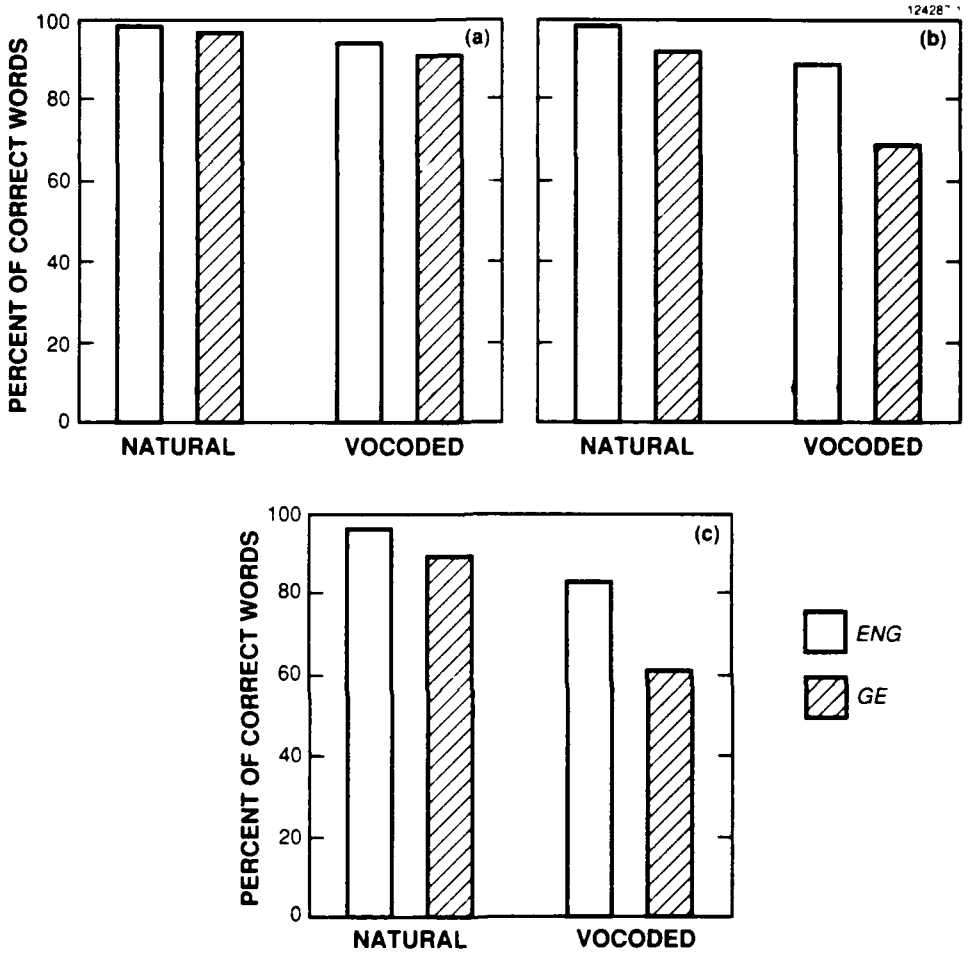


Figure 2-1. Percent of correct words for each of the three intelligibility tests: (a) Diagnostic Rhyme Test (DRT), (b) Meaningful Sentences Test (MST), and (c) Semantically Anomalous Sentences Test (SAST).

TABLE 2-1.

Diagnostic Rhyme Test: Number of Erroneous Words*

Natural Speech				LPC-Vocoded Speech			
English Natives		German-English Bilinguals		English Natives		German-English Bilinguals	
BH	1	AR	4	AS	4	CM	23
CQ	3	CB	1	BS	11	CW	18
DW	4	EE	4	CR	7	CZ	22
GM	1	EH	2	DM	7	DD	22
JL	3	EM	11	JG	9	HK	16
KC	1	HE	3	PE	15	JB	20
LH	1	HH	2	RG	24	KS	16
ML	3	HS	9	KB	14	SM	13
SG	1	KB	3	TW	19	SS	13
SZ	2	SB	3	WS	13	UT	14
Sum	20		42		123		177
$\bar{x}$	2.0		4.2		12.3		17.7
Std Dev	1.2		3.2		6.1		3.9

*Maximum possible errors per subject = 232

Scores for the MST were lower than for the DRT [Figure 2-1(b)]. The percent of correctly rendered words for each group was as follows: ENG-NAT = 99.03 percent, GE-NAT = 92.09 percent, ENG-VOC = 89.74 percent, and GE-VOC = 69.71 percent. (Note that, for the ENG-NAT subjects, the MST mean percentage was only a fraction of a percent lower than their uncorrected DRT percentage while, for all other groups, the MST mean percentage was about 5 to 22 points lower.) Raw error data (Table 2-2) revealed even larger between- and within-group differences than suggested by the percentages. For example, while the GE-NAT group's mean MST score was only about 7 percentage points lower than the ENG-NAT subjects', their actual errors (as reflected in the raw data) exceeded those of the ENG-NAT group by a factor of over 8.

On the SAST, scores were lower than on the DRT or the MST [Figure 2-1(c)]. The percentages for each group were as follows: ENG-NAT = 97.42 percent, GE-NAT = 90.18 percent, ENG-VOC = 83.52 percent, and GE-VOC = 62.11 percent. On the average, subjects scored about 2 to 8 percentage points lower on the SAST than on the MST. Furthermore, for the ENG-NAT group, the percent correct was still quite high (over 97 percent). Analysis of the raw error data (Table 2-3) also revealed that the GE-NAT group had nearly four times as many errors as the ENG-NAT group, while the GE-VOC group had over twice as many errors as the ENG-VOC group.

A three-way repeated-measures ANOVA with two-between and one-within subjects factors was conducted on the sentence data. It revealed a significant main effect for language group [F (1, 36)

TABLE 2-2.

Meaningful Sentences Test: Number of Erroneous Words*

Natural Speech				LPC-Vocoded Speech			
English Natives		German-English Bilinguals		English Natives		German-English Bilinguals	
BH	1	AR	35	AS	42	CM	72
CQ	3	CB	8	BS	46	CW	145
DW	3	EE	17	CR	51	CZ	105
GM	1	EH	16	DM	30	DD	179
JL	1	EM	45	JG	42	HK	111
KC	5	HE	43	PE	34	JB	138
LH	19	HH	56	RG	55	KS	133
ML	1	HS	55	TK	37	SM	121
SG	2	KB	17	TW	22	SS	71
SZ	1	SB	11	WS	34	UT	85
Sum	37		303		393		1160
$\bar{x}$	3.7		30.3		39.3		116.0
Std Dev	5.5		18.6		9.9		34.4

*Maximum possible errors per subject = 383

= 68.95, $p < 0.0001$], a significant main effect for listening condition [F (1, 36) = 120.85, $p < 0.0001$], and a significant main effect for sentence type [F (1,36) = 93.14, $p < 0.0001$]. There was also a significant language-group × listening-condition interaction [F (1,36) = 16.56, $p < 0.0002$], and a significant listening-condition × sentence-type interaction [F (1, 36) = 32.80, $p < 0.0001$].

A Tukey post hoc test revealed that the mean number of errors made on the MST by the ENG-NAT group was significantly smaller than that made by the GE-NAT group ($p < 0.05$); the mean number of errors made on the MST by the ENG-VOC group was significantly smaller than that made by the GE-VOC group ($p < 0.01$); and both the native and non-native listeners had significantly more errors on the MST in the vocoded than in the natural condition ($p < 0.01$). Post hoc analysis of SAST errors revealed a pattern of significance identical to that obtained for the MST.

2.5.1.2 Erroneous Sentences

Further analysis of the data included a tabulation of the percent of correct MST and SAST sentences—i.e., sentences rendered correctly in their entirety. Although this measure of performance is not independent of the percentage of words rendered correctly, it was believed that it could provide

TABLE 2-3.

Semantically Anomalous Sentences Test: Number of Erroneous Words*

Natural Speech				LPC-Vocoded Speech			
English Natives		German-English Bilinguals		English Natives		German-English Bilinguals	
BH	4	AR	64	AS	66	CM	92
CQ	8	CB	16	BS	81	CW	171
DW	6	EE	12	CR	65	CZ	152
GM	3	EH	18	DM	44	DD	184
JL	1	EM	49	JG	59	HK	147
KC	18	HE	62	PE	50	JB	172
LH	36	HH	38	RG	100	KS	168
ML	2	HS	60	TK	57	SM	155
SG	6	KB	35	TW	59	SS	105
SZ	15	SB	22	WS	50	UT	105
Sum	99		376		631		1451
$\bar{x}$	9.9		37.6		63.1		145.1
Std Dev	10.7		20.2		16.6		32.7

*Maximum possible errors per subject = 383

additional important information about the subjects' intelligibility performance since it entailed a more rigorous criterion of correctness.

On the MST, the ENG-NAT group rendered correctly an average of 94.74 percent of the 57 sentences. For the GE-NAT group, the corresponding value was 64.56 percent. The ENG-VOC group rendered an average of 53.86 percent of the sentences correctly, compared with 20.18 percent for the GE-VOC group [Figure 2-2(a)]. In terms of raw error data, the ENG-NAT group had a mean of only three erroneous sentences, while the GE-NAT group had 20.2. The ENG-VOC group had a mean of 26.3, and the GE-VOC had a mean of 45.5.

Even larger between-group differences emerged in the scores for the SAST [Figure 2-2(b)]. Here the ENG-NAT group rendered 87.54 percent of the sentences correctly, while the GE-NAT group rendered 56.14 percent correctly. The ENG-VOC group rendered 30.35 percent correctly, and the GE-VOC rendered 6.67 percent correctly. In terms of raw errors, the ENG-NAT group had a mean of 7.1, the GE-NAT group had a mean of 25.0, the ENG-VOC group had a mean of 39.7, and the GE-VOC group had a mean of 53.2. Thus, on the average, the non-native listeners had fewer than four correct sentences (out of a possible total of 57) on the SAST.

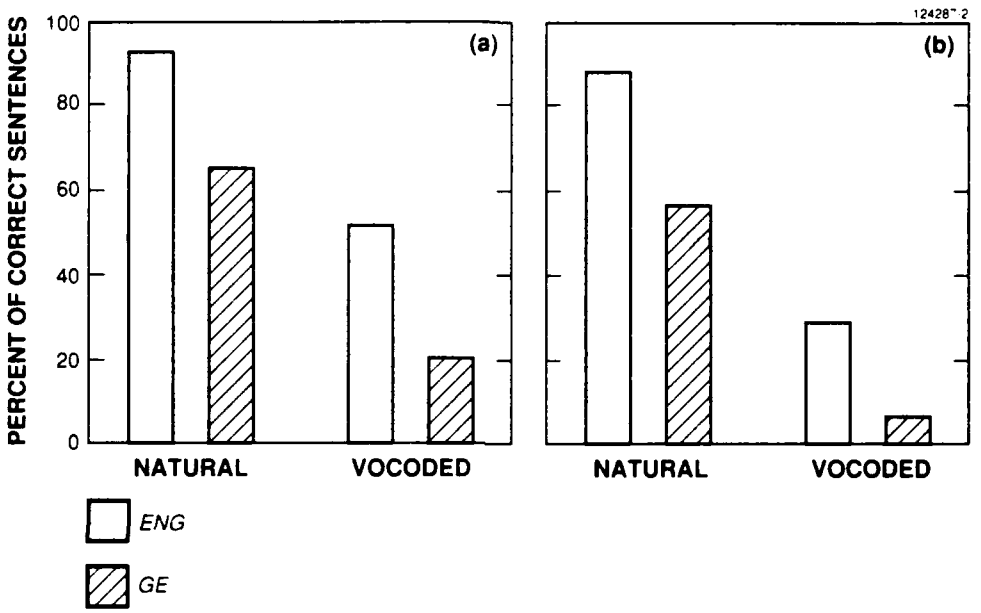


Figure 2-2. The percent of correct sentences (i.e., sentences in which all *words were rendered correctly): (a) MST and (b) SAST.*

2.5.1.3 Errors and the Independent Variables

Also calculated were the differences in scores on the MST and SAST for both language groups and listening conditions. This analysis revealed that the effects of the three independent variables (native language, listening condition, and test type) were nonadditive. This is apparent if the mean number of errors made by the ENG-NAT group on the MST (3.7) is treated as a baseline value to which the values associated with the other variables (and combinations of) are compared—as is done in Table 2-4. (Values in the third column represent the difference between the specified group's mean number of errors and those produced by the ENG-NAT group on the MST.)

TABLE 2-4.

Mean Number of Errors and Difference Scores

Group	Test	Mean Number of Errors	Difference
ENG-NAT	SAST	9.9	6.2
GE-NAT	MST	30.3	26.6
GE-NAT	SAST	37.6	33.9
ENG-VOC	MST	39.3	35.6
ENG-VOC	SAST	63.1	59.4
GE-VOC	MST	116.0	112.3
GE-VOC	SAST	145.1	141.4

If a simple additive relationship existed among the variables, it would be predicted that 6.2 errors should be added to the ENG-NAT MST baseline (3.7) for the SAST, 26.6 errors should be added for the non-native listeners, and 35.6 errors should be added for vocoding. Thus, the ENG-VOC group should have had a mean of 39.3 errors on the SAST (3.7 + 35.6), but they had 63.1; the GE-VOC group should have had a mean of 65.9 errors on the MST (3.7 + 26.6 + 35.6), but they had 116.0; and the GE-VOC group should have had a mean of 72.1 errors on the SAST (3.7 + 6.2 + 26.6 + 35.6), but they had 145.1.

However, there is some evidence of a constant differential (k) for ENG versus GE groups regardless of the test method. This constant does, however, depend upon speech quality (Figure 2-3).

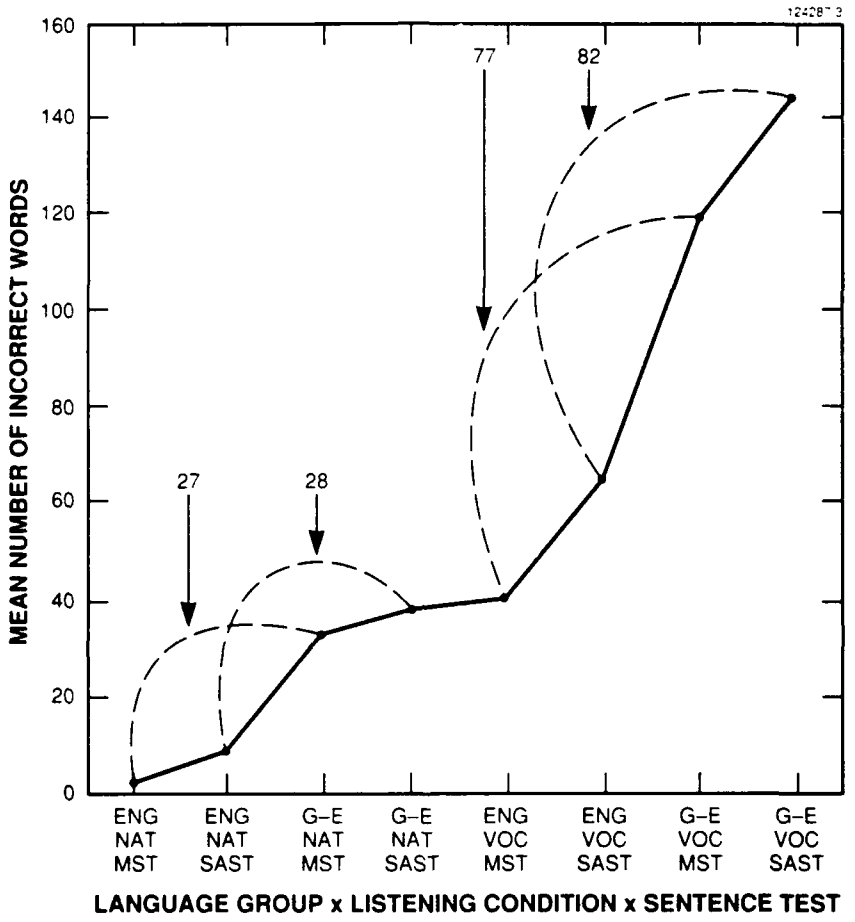


Figure 2-3. Mean number of incorrect words categorized by language group, listening condition, and sentence test. (The difference between each pair of scores compared appears above the arrow denoting the comparison.)

That is, in the natural-speech condition, there was a mean difference of about 27 errors between the ENG and GE groups on *both* sentence tests. Moreover, in the vocoded-speech condition, there was a mean difference of about 80 errors between the ENG and GE groups on *both* sentence tests. Thus, this analysis suggests the presence of two constants—k_1 and k_2—one for natural and one for vocoded speech. [6]

2.5.2 Other Types of Errors

2.5.2.1 Position-in-Sentence Errors

This analysis revealed some systematicity in the between-group error patterns associated with each one of the intrasentential word positions [Figures 2-4(a) and (b)]. For example, subject groups produced relatively few errors in position 1 (either an article or null) and, overall, relatively few errors in position 5 (also an article). Subjects also produced, in general, more errors in position 5 than position 1 (possibly due, at least in part, to the larger number of articles in position 5 than 1). Further, as examination of proportional data shows (Tables 2-5 and 2-6), for nearly all groups the word in position 4 (the verb) tended to have the largest proportion of errors. In addition, for all groups, a smaller proportion of errors was associated with words in sentence-final position in the MST than in the SAST.

However, some differences in the performance of the native and non-native subjects did emerge. For example, on the MST, the GE-VOC group exhibited a relatively low proportion of errors (0.13) on words in position 5 (the determiner) while the ENG-VOC group exhibited an extremely high proportion of errors (0.35) in this position. Moreover, with only one exception (the SAST in the natural-speech condition), the native subjects had a smaller proportion of errors in positions 2 and 3 (adjective and noun) than did the non-native subjects.

2.5.2.2 Omissions

Analysis of the number of words omitted on the sentence tests (no responses were omitted on the DRT) revealed that the ENG-NAT and ENG-VOC groups omitted an average of five or fewer words on the MST and seven or fewer on the SAST [Figures 2-5(a) and (b)]. The GE-NAT group likewise omitted very few words on the MST or SAST. However, the GE-VOC subjects omitted an average of 50.8 words (about 13 percent) on the MST and 30.1 (about 8 percent) on the SAST.

A three-way repeated-measures ANOVA with two between- and one-within subjects factors revealed a significant main effect for language [$F(1,36) = 20.51$, $p < 0.0005$], listening condition

[6] Clearly, the values obtained for k_1 and k_2 are dependent upon which variables are held constant and which are changed when difference scores are obtained. For example, different values of k would result if all variables were held constant except test type (e.g., ENG-NAT MST versus ENG-NAT SAST) rather than language background (e.g., ENG-NAT MST versus GE-NAT MST). This can be demonstrated easily if the first four data points in Figure 2-3 are designated (from left to right) A, B, C, D. If C - A = D - B, then B - A = D - C.

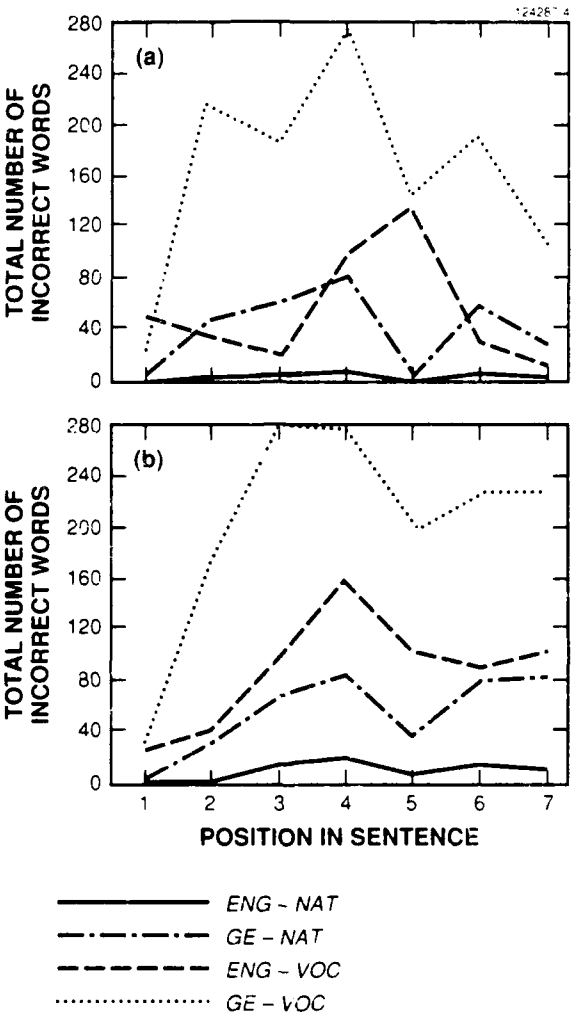


Figure 2-4. Total number of errors associated with each of the seven intrasentential word positions for the: (a) MST and (b) SAST.

TABLE 2-5.

Meaningful Sentences Test: Distribution of Positional Errors*

	1 (Art)	2 (Adj)	3 (Noun)	4 (Verb)	5 (Art)	6 (Adj)	7 (Noun)
Natural							
English Natives	0.00	0.05	0.11	0.35	0.03	0.35	0.11
G-E Bilinguals	0.01	0.16	0.20	0.28	0.01	0.22	0.11
Vocoded							
English Natives	0.12	0.09	0.06	0.25	0.35	0.08	0.04
G-E Bilinguals	0.02	0.19	0.16	0.24	0.13	0.16	0.09

*Values are proportions of total number of errors for each group.
Row totals may not equal 1.0 due to rounding errors.

TABLE 2-6.

Semantically Anomalous Sentences Test: Distribution of Positional Errors

	1 (Art)	2 (Adj)	3 (Noun)	4 (Verb)	5 (Art)	6 (Adj)	7 (Noun)
Natural							
English Natives	0.01	0.03	0.18	0.23	0.15	0.21	0.19
G-E Bilinguals	0.01	0.09	0.18	0.22	0.10	0.20	0.21
Vocoded							
English Natives	0.04	0.06	0.15	0.25	0.17	0.15	0.17
G-E Bilinguals	0.02	0.13	0.19	0.19	0.14	0.16	0.16

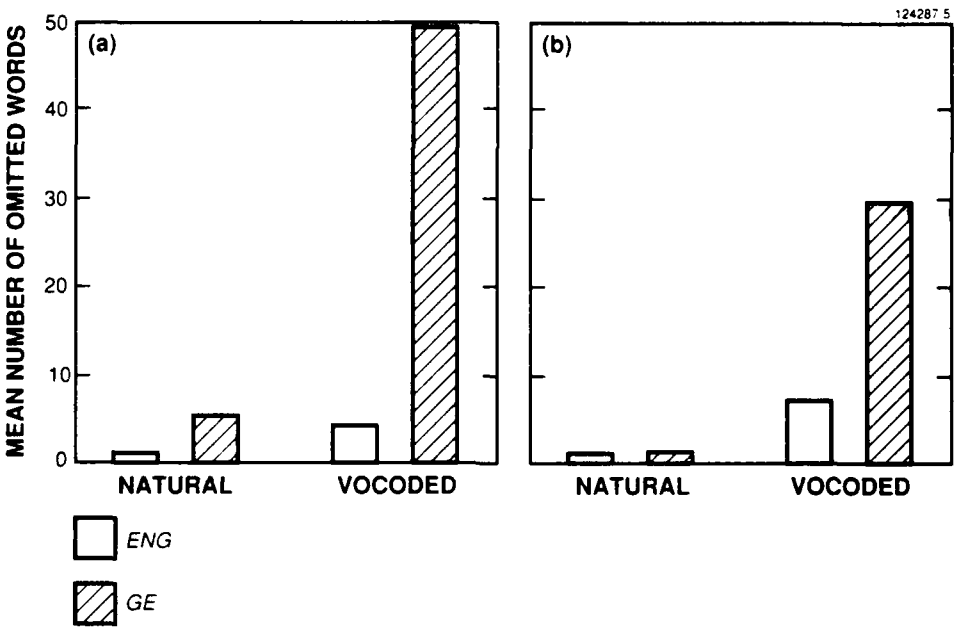


Figure 2-5. Mean number of words omitted on the: (a) MST and (b) SAST.

[F (1,36) = 26.46, $p < 0.0001$] and sentence type [F (1,36) = 10.901, $p < 0.0025$]. All interactions were significant at the 0.01 level. Pairwise comparisons of the number of omission errors were conducted to determine whether or not there was a difference in the number of omissions made on the MST and SAST. Results revealed no significant difference for the ENG-NAT group, but a significant difference for the ENG-VOC group: this group had significantly more omissions on the SAST than on the MST ($p < 0.01$). The opposite result obtained for the two GE groups—i.e., the GE-NAT group had significantly *fewer* omissions on the SAST than on the MST ($p < 0.05$), as did the GE-VOC group ($p < 0.01$).

2.5.2.3 Frequency-Based Errors

The DRT data yielded somewhat inconsistent results with respect to the number of errors associated with words of low, mid, and high frequency [Figure 2-6(a)]. Although a three-way repeated-measures ANOVA revealed a significant main effect for frequency [F (2,72) = 3.34, $p < 0.05$], there were no significant interactions.

On the MST and SAST, a somewhat different pattern of word-frequency effects emerged. A four-way repeated-measures ANOVA yielded four significant main effects, including a significant main effect for frequency [F (2,72) = 150.17, $p < 0.0001$]. All but one of the interactions (language group × speech condition × sentence type) was significant, including a significant language-group × listening-condition × sentence-type × word-frequency interaction [F (2,72) = 4.48, $p < 0.02$].

Post hoc analyses revealed that, in the sentence tests, vocoded speech resulted in more significant word-frequency effects than did natural speech, and the MST was associated with more frequency-based differences than the SAST. In addition, the native subjects showed fewer significant word-frequency effects than did the non-native subjects. That is, while three of the pairwise comparisons were significant for the ENG subjects, ten were significant for the GE subjects.

2.5.2.4 Sentence-Order Errors

In order to evaluate the possibility that subjects' performance improved with practice over the course of a single test, the number of errors made on the first 15 sentences in the sentence tests was compared with the number of errors made on the last 15 sentences. (No such analysis was carried out with the DRT data because there were too few errors to make a meaningful comparison.)

Results indicated that, on both the MST and SAST, ENG-NAT subjects made nearly equal numbers of errors on the first and last 15 sentences. Such a result was *not* observed for the GE-NAT subjects [Figure 2-7(a)]. In the vocoded condition, differences in the error patterns of the native and non-native groups also emerged [Figure 2-7(b)].

A four-way repeated-measures ANOVA conducted on the sentence data revealed four significant main effects, including a significant main effect for sentence-order errors (errors on the first versus last 15 sentences) [F (1,36) = 24.87, $p < 0.0001$]. One of the interactions (language group ×

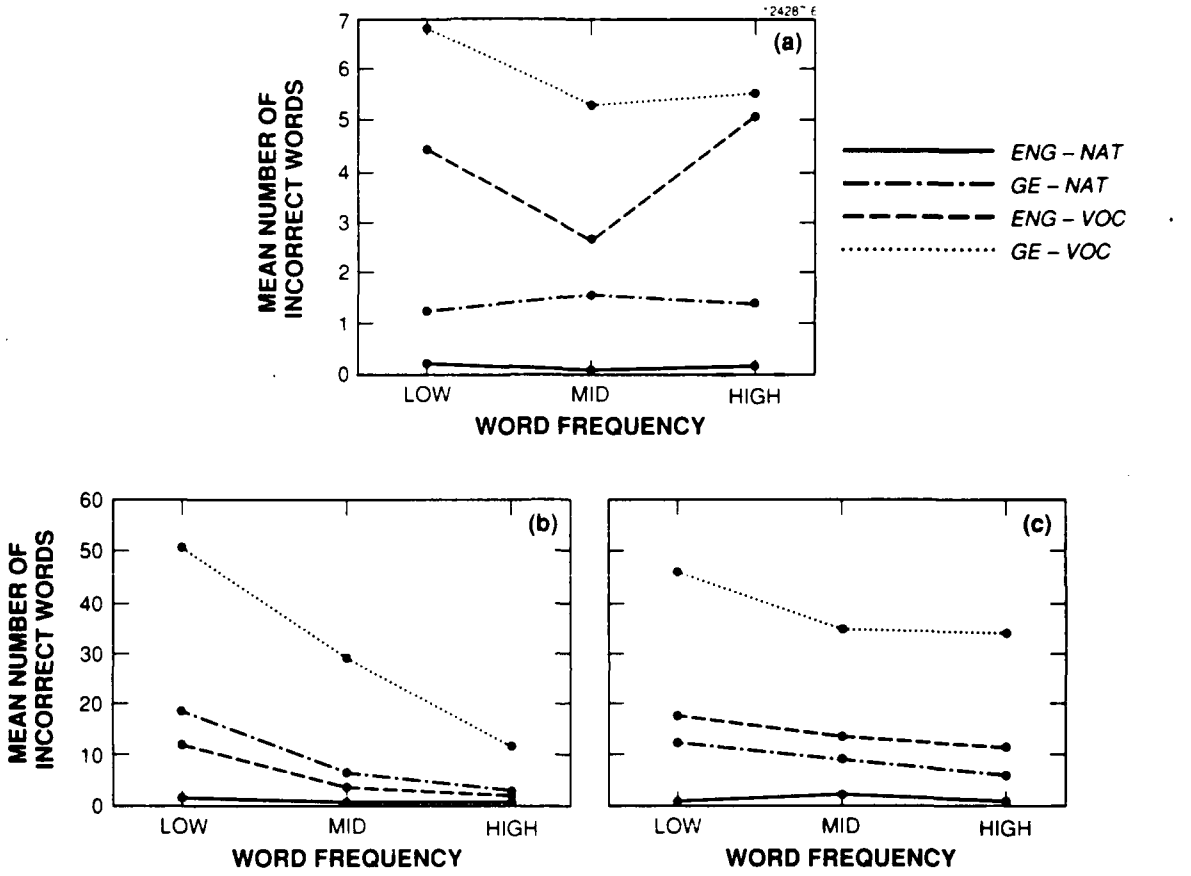


Figure 2-6. Mean number of incorrect words on the three tests classified according to word frequency: (a) DRT, (b) MST, and (c) SAST.

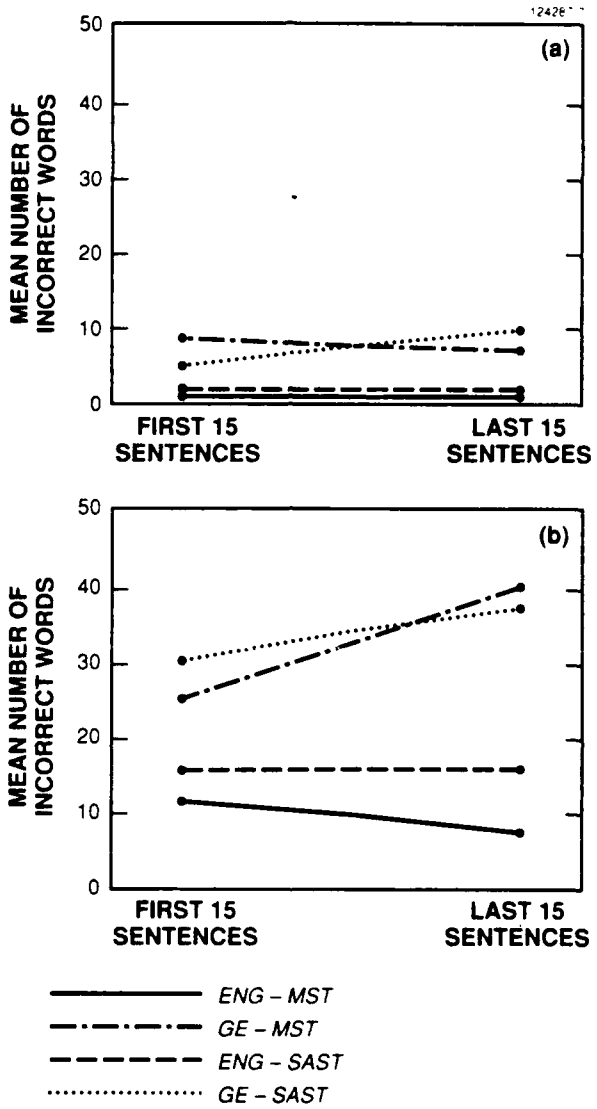


Figure 2-7. Mean number of words rendered incorrectly in the first 15 and last 15 sentences on the MST and SAST: (a) natural speech and (b) vocoded speech.

sentence type × error order) was not significant, but all others were, including a language-group × listening-condition × sentence-type × error-order interaction [$F(1, 36) = 19.58$, $p < 0.001$].

Pairwise comparisons conducted on the error-order data revealed that the GE-NAT group had significantly *fewer* errors on the first than on the last 15 sentences in the SAST. The GE-VOC group also exhibited significantly fewer errors on the first than on the last 15 sentences—on both the MST and the SAST.

3. DISCUSSION

It will be recalled that the objectives of the present study were to address the following questions: (1) To what extent is intelligibility affected when vocoded versus natural stimuli are presented to native and non-native listeners? (2) What is the magnitude of between-test differences when various types of intelligibility tests are utilized with these two groups? (3) Are specific patterns of response similar for native and non-native listeners? These questions can now be addressed in light of the results obtained in this study.

3.1 Overall Errors

3.1.1 Effect of Vocoding

Findings clearly revealed the deleterious effect of LPC-vocoded speech upon intelligibility. When averaged across all three test types (DRT, MST, and SAST), the percent correct for the native listeners presented with vocoded speech was about 88 percent, compared with over 98 percent for natural speech. Likewise, for the non-native listeners, vocoding resulted in an overall average of about 72 percent correct, compared with 93 percent for natural speech.

In addition, there was an interaction between listening condition and language background. For the native listeners, vocoding reduced the percent correct by about 14 percentage points on the SAST. But for the non-native listeners, vocoding reduced the percent correct by about 28 points on the SAST.

These findings are not in complete agreement with those previously obtained by Mack and Tierney (1987) who observed a larger relative difference in the natural and vocoded conditions (on the SAST) for the natives than for the non-natives as may be seen in Table 3-1. The difference in the results of the two experiments may be due to the fact that Mack and Tierney used high-quality channel-vocoded speech while the present study used LPC-vocoded speech of somewhat lesser quality (i.e., it was more degraded acoustically than the channel-vocoded stimuli). In fact, the SAST performance of the natives presented with channel-vocoded speech in the Mack and Tierney study was over 8 percentage points higher than that of the natives presented with LPC-vocoded speech in the present study. This finding demonstrates that there were fairly robust differences in the quality of the coding systems used in the two studies, and it suggests that the difference in the intelligibility of natural and vocoded speech—as reflected in percent correct—is greater for non-natives than natives only if the speech is of relatively low quality.

3.1.2 Effect of Test Type

Also of interest in the overall error analysis was the fact that subjects obtained, on the average, very high scores on the DRT. Moreover, in terms of their uncorrected scores, they scored better on

TABLE 3-1.

Comparison of Scores from Two Studies*

Mack and Tierney (1987): SAST Scores		
	English Natives	German-English Bilinguals
Natural	98.72	85.29
Vocoded	92.01	81.38
Difference	6.71	3.91
Mack, Tierney, and Boyle (1990) SAST Scores		
	English Natives	German-English Bilinguals
Natural	97.42	90.18
Vocoded	83.52	62.11
Difference	13.90	28.07

*All values are in percentages.

the DRT than on the MST, and better on the MST than on the SAST. These findings raise two important issues of relevance to intelligibility research.

First, if high-quality coded speech is used in testing, the DRT may not be sufficiently sensitive due to (possibly) subtle differences between various speech systems. That is, if a "ceiling effect" emerges on the DRT such that most subjects obtain scores near 100 percent, then there is a restricted range of scores and a skewed distribution. Hence, it becomes difficult to make valid comparisons of different systems. Second, due to the fact that the DRT places fewer perceptual, cognitive, and recall demands upon subjects, it may yield scores for native and non-native listeners that are quite similar—as was observed in the present study.

Thus, while possibly providing important insight about the performance of listeners presented with a restricted and highly familiar range of lexical options, the DRT may not provide valid information about performance in more demanding communication contexts.

Of course, the extent to which the results of any intelligibility test can be generalized to "real world" communication cannot be completely understood. For some applications—as in messages conveyed to and relayed by pilots in a cockpit—the DRT may be an appropriate diagnostic instrument, since the context to which its results are generalized is one in which listeners utilize a constrained and highly familiar set of lexical items and, possibly, a restricted set of syntactic structures. However, for other applications—as in the text-to-speech conversion systems designed for the handicapped—the DRT may not be an appropriate diagnostic. Here the context is one in which speech is not highly restricted: thousands of lexical items and numerous syntactic structures may be used. What seems reasonable, then, is that results of different intelligibility tests be compared in order to provide information about levels of performance as a function of various stimulus types and task demands.

Another point to be addressed with respect to test type is that the performance of all subject groups on the MST was superior to performance on the SAST. However, the difference between the MST and SAST scores for the non-native listeners was proportionately smaller than the difference for the native listeners. (In both the natural and vocoded conditions, the non-natives had only about 1.25 as many errors on the SAST as on the MST, while the natives had 1.6 to 2.7 times as many). Apparently, the meaningful sentences were facilitative for the native listeners because these sentences were more predictable than the anomalous sentences. Thus the native listeners' knowledge of English (or their ability to utilize that knowledge in the intelligibility task) was superior to the non-natives'. The meaningful sentences may not have been as facilitative for the non-native listeners because the sentences were not sufficiently predictable. (It will be recalled that, on the MST, not all sentences were highly predictable.) Thus, many of the meaningful sentences may have been treated as anomalous by the non-natives.

Yet the fact that both groups performed at least somewhat better on the MST than on the SAST is consistent with related findings of previous studies demonstrating that stimuli of high predictability are responded to more accurately or more rapidly than stimuli of low predictability (Kalikow, Stevens, and Elliot, 1977; Salasoo and Pisoni, 1985; Schmidt-Nielsen and Kallman, 1987; Boothroyd and Nittrouer, 1988). Indeed, Salasoo and Pisoni (1985) state that "meaningful sentence contexts support faster, more efficient, and qualitatively different identification processes than semantically anomalous sentence contexts" (p. 221). What is also evident is that the extent to which predictability influences intelligibility may be determined by the language experience of the listener.

3.1.3 Effect of Language Background

There were large and consistent differences in the performance of the native and non-native listeners in the present study, with the non-natives exhibiting more errors than the natives on all tests and in all listening conditions—with only one exception: there was no significant difference in the number of errors produced on the DRT in the natural condition by the native and non-native listeners.

The mean number of erroneous words produced in response to the SAST by the non-natives in the vocoded condition was especially large (145.1 or over 37 percent incorrect). And when sentences rather than words were analyzed, it was found that the GE-VOC group rendered an average of about 12 of the MST sentences and fewer than four of the SAST sentences correctly in their entirety. (The corresponding values for the ENG-VOC group were about 31 and 17.) Even when stimuli were not vocoded, the non-native listeners had fairly low intelligibility scores. For example, on the MST, they had an average of over eight times as many errors as the native listeners. These findings revealed that, on an intelligibility test that places considerable demands upon mechanisms of sentence processing and recall, even highly fluent non-native listeners may perform quite poorly. This is especially interesting in view of the fact that all of the non-natives in the present experiment had begun their study of and/or exposure to English at a relatively early age and all were, at the time of testing, immersed in an English-language milieu.

There are at least three possible explanations for the relatively poor performance of the non-native listeners. One is that they were simply less proficient in processing English than the natives because their exposure to English had not been as extensive. Another is that their internalized English system was fundamentally different from the natives', possibly as a result of differences in the groups' manner of English acquisition and/or differences in the age at the onset of English acquisition. A third possible explanation is that the poorer performance of the non-natives was a result of the transfer of word- and/or sentence-processing strategies from their native German. Such strategies could have been unproductive in an English-language task. It is also possible, of course, that the non-natives' intelligibility performance was due to a combination of these three factors.

Whatever the cause underlying the non-natives' performance, the results of the present study suggest that, in normal communication, listeners such as these either rely heavily upon contextual cues for comprehension *or* that they ordinarily experience relatively serious problems in comprehension.

3.1.4 Effect of Interaction Among Variables

A final important aspect of the overall error analysis was the finding that the relationships among the independent variables were nonadditive. That is, when the mean number of errors made on the MST by the ENG-NAT group was treated as a baseline to which error rates for the other tests, conditions, and groups were compared, it was found that a simple additive relationship did not account for the observed error data. If it had, the GE-VOC group would have made an average of about 72 errors on the SAST, but they had an average of 145.

Nonetheless, some systematicity was observed when two values of k were derived as follows:

$$k_1 = \bar{x} \text{ errors GE} - \text{NAT} - \bar{x} \text{ errors ENG} - \text{NAT}$$

$$k_2 = \bar{x} \text{ errors GE} - \text{VOC} - \bar{x} \text{ errors ENG} - \text{VOC}$$

It was found that, in the present study, $k_1 = 27$ and $k_2 = 80$.

However, it seems desirable to attempt to discover other quantitative approaches that can accurately reflect and predict the absolute or relative decrement in intelligibility as a function of test type, listening condition, and language experience. For example, Boothroyd and Nittrouer (1988) present an equation for obtaining the value of k that is equal to the ratio of the probabilities of recognition of speech units with and without context:

$$k = log(1 - p_c)/log(1 - p_i)$$

where p_c = probability of recognizing a speech unit in context and p_i = probability of recognizing a speech unit without context (i.e., in isolation or in nonsense material).

If a logarithmic relationship existed between the values obtained in the present study, the value of k should have been equal for all groups and conditions (if k can be obtained by dividing the

logarithm of the probability of recognizing words in the meaningful sentences by the logarithm of the probability of recognizing words in the anomalous sentences). However, this was not found. Although k was nearly the same for the ENG-NAT and GE-NAT groups (0.9636 and 0.9570, respectively), it was different for the ENG-VOC and GE-VOC groups (0.8666 and 0.8399, respectively). It also differed fairly substantially for the ENG-NAT and GE-NAT groups.

3.2 Other Types of Errors

3.2.1 Differences in Processing Strategies

In recent years, much attention has been given to processing strategies and to characterizing the conditions under which the use of different strategies is encouraged. One of the major distinctions made has involved the difference between top-down (theory-driven) and bottom-up (data-driven) strategies. Marslen-Wilson and Welsh (1978) have stated that a top-down strategy employs "higher-level constraints ... to 'drive' the processing of the input data," while a bottom-up strategy utilizes the specific properties of the data to determine the "higher-level representation" of the input (p. 30). This characterization of the two processes has been reformulated by some researchers into a distinction between utilizing lexico-semantic and pragmatic knowledge to understand language input and utilizing the phonetic and morpho-syntactic properties of the input in order to arrive at a meaningful representation (e.g., Cziko, 1980). Although, as Mack (1988) has pointed out, there may be no simple mapping between linguistic components, types of errors, and processing strategies, it may still be possible to gain some insight into the top-down/bottom-up distinction by examining the data obtained in the present study. In fact, some evidence of between-group differences in processing strategies may be discerned in the position-in-sentence error data.

Specifically, the position-in-sentence data revealed that the distribution of errors in the natural and vocoded conditions was fairly similar with one notable exception: For the ENG-VOC group, 0.35 of the errors made were in position 5 (the article). This value was *far* greater than that exhibited by the ENG-NAT group (0.03), and it may have reflected inattention to the precise form of the article or an inability to hold the correct form in memory. This apparent difficulty with the article suggests that the ENG-VOC group was utilizing a predominantly top-down (theory-driven) processing strategy that led them to attend to the lexico-semantic aspects of the stimuli. By contrast, the GE-NAT and GE-VOC groups may have utilized a bottom-up (data-driven) approach whereby they attended to the detailed acoustic/phonetic features of the stimuli. This interpretation is consistent with the conclusion of other researchers (Cziko, 1980; McLeod and McLaughlin, 1986) who have concluded that non-natives utilize bottom-up processing strategies to a greater extent than do natives.

3.2.2 Difficulty in Word Recall

In tests such as the MST and SAST, moderate to severe demands are placed upon short-term memory. In fact, it could be maintained that, because subjects in the present study were required

to transcribe six- and seven-word sentences on the MST and SAST, their ability to recall was tested just as their intelligibility performance was. This is no doubt true, but it does not render the recall and transcription-based approach invalid as a test of intelligibility. (Indeed, even in normal sentence processing, memory is sometimes strongly implicated.) What is of interest is that, in addition to providing evidence of intelligibility problems, the sentence tests also provided evidence of recall difficulty.

On both the MST and SAST, the GE-VOC group omitted *far* more words than did the ENG-VOC or the GE-NAT groups. It seems that, for this group, vocoding induced especially serious recall problems. In discussing decrements in subjects' recall of synthetic words and digits, Luce, Feustel, and Pisoni (1983) state that synthetic speech places "increased demands on encoding and/or rehearsal processes in short-term memory" (p. 28). Thus, it seems that the problems of the GE-VOC group were compounded: not only were they presented with vocoded speech, but they lacked nativelike proficiency in English. Together, these factors resulted in serious intelligibility problems.

It could of course be argued that the GE-VOC subjects omitted words not because they forgot them, but because they simply failed to recognize them (i.e., they did not know them) or because acoustic distortion rendered the words unrecognizable. The former explanation seems unlikely because these subjects omitted even quite common words; but the latter interpretation is possible. It may be that non-native listeners require highly accurate and complete (even redundant) feature specification if they are to recognize speech stimuli correctly. If features are distorted or absent—as they may be when speech is vocoded—then non-native listeners may be unable to process those portions of the stimulus that remain fairly well specified; hence they cannot recognize a word at all. Thus, the disproportionately high number of omissions found in the GE-VOC group may have been due to serious problems in recall caused by capacity "overload" induced by acoustically degraded speech—or their omissions may have been due to problems in encoding resulting from incorrect or incomplete representation of acoustic/phonetic features in the stimuli, rendering the non-natives unable to utilize other available (intact) acoustic/phonetic cues.

3.2.3 Word-Frequency Effects

Previous studies have found that word frequency may affect speech intelligibility rather substantially. In the present study, significant word-frequency effects were also observed, although the effects varied as a function of test type, listening condition, and language group.

Overall, the MST showed the largest number of significant differences in words of high, mid, and low frequency, and vocoded speech yielded more differences than natural speech. Moreover, on the sentence tests, there were more significant differences in words of different frequency for the non-native than for the native listeners.

It is not readily apparent why the MST showed stronger word-frequency effects than the SAST. It may be that the MST sentences were processed "naturally"—with subjects able to utilize linguistic knowledge so that words of high frequency (i.e., more common words) were more

accurately perceived and or recalled. On the other hand, if subjects treated the anomalous sentences "unnaturally"—not as sentences but as collocations of unrelated words (Salasoo and Pisoni, 1985)—then they may have been less able to utilize linguistic knowledge to facilitate processing or recall.

Yet such an explanation is difficult to accept, given that the non-native listeners exhibited stronger word-frequency affects than did the natives. (At least they showed a higher number of significant differences between words of different frequencies.) It is thus possible that the effect of word frequency is most salient when the intelligibility task utilizes the types of mechanisms used in ordinary communication (as the MST might). If stimuli are isolated words, or if they can be treated as such, then the word-frequency effect is diminished.

Further, that the non-natives exhibited *any* significant word-frequency effects seems especially strong support for the notion that they were sensitive to the frequency of words in their non-native language. That they were apparently more sensitive to word-frequency effects than the native listeners suggests that they may have found words of mid and low frequency more unusual than the natives did and thus more difficult to process. This interpretation finds indirect support in a study by Mullenix, Pisoni, and Martin (1989) who note that "word frequency manipulations produce greater effects when the information specifying the items is ambiguous or degraded" (p. 375)—i.e., when the listening situation is less than ideal. Thus it may be possible to interpret a lack of nativelike competence as one more factor that reduces speech intelligibility and thereby contributes to word-frequency effects.

3.2.4 Perceptual Learning Versus Test Fatigue

It was deemed important to explore the possibility that, over the course of a single test, subjects' performance might improve. This was of special interest in the case of the vocoded stimuli for, if significant improvement in intelligibility was observed within a given test, it could have major implications for perceptual learning (and for the possible efficacy of speech-systems training).

Analyses comparing the number of errors made on the first 15 and last 15 sentences in the MST and SAST revealed that, on the SAST for the GE-NAT group, and on the MST and SAST for the GE-VOC group, subjects had significantly *more* errors on the last than on the first 15 sentences, suggesting the effects of test fatigue.

It appears that under conditions of only moderate difficulty, intelligibility scores may remain constant (or may even improve) over a single test. But when task conditions are especially demanding—as they apparently were for the GE-NAT group on the SAST and for the GE-VOC group on the MST and SAST–a decrement in performance may occur. This conclusion is supported by a study by Clark, Dermody, and Palethorpe (1985) who found that intelligibility performance as a function of stimulus repetition improved to a greater extent when natural rather than synthetic (MITalk) speech was used.

Thus it seems that non-native listeners are at a particular disadvantage when confronted with a task in which they must encode and then reproduce semantically complex and/or acoustically degraded speech. Not only does their performance fail to improve over the course of a test, but it actually worsens, possibly reflecting the effects of fatigue, or even of frustration with the difficulty of the task and a consequent inattention to the stimuli. These findings also suggest that it may be fairly easy to overload the non-native processing system. Hence, non-natives may require more "rest" than natives when they are listening to coded speech. At the least, these results reveal that repeated exposure to acoustically degraded stimuli does not have the same effect upon non-native as upon native listeners.

4. SUGGESTIONS FOR FUTURE RESEARCH

The possible directions for future research on speech intelligibility suggested by the present study are myriad. Yet there are several avenues of work that could prove particularly fruitful.

First, it is necessary to determine the extent to which the performance of the subjects in the present study can be generalized to other types of listeners. To do so requires that additional native and non-native listener groups be tested. Of related interest is the performance of native listeners of English whose dialects differ, perhaps markedly, from that of the producer of the test stimuli. (For example, would such listeners show any decrement in performance or would their native-language competence enable them to process accurately stimuli presented in a "non-native" dialect?)

Second, continued attempts should be made to compare the *patterns* of response on intelligibility tests made by members of different language groups. Most obvious would be detailed cross-linguistic analyses of phonological errors. But other types of errors, such as position-in-sentence errors or word omissions, may be of special interest if it can be demonstrated that they reveal differences in processing strategies and recall abilities. An analytic procedure such as response coincident analysis (Baker, Hogan, and Rozsypal, 1988) might also prove effective in distinguishing specific types of listeners—even those within same-language groups.

Third, it is important to determine which test is needed for the assessment of intelligibility in specific types of contexts. For example, as indicated above, the DRT may be the most appropriate test for evaluating speech systems to be used in certain communication contexts, while the MST or SAST may be more appropriate for evaluating systems to be used in other contexts. It is clear that continued research, using a variety of tests, is needed in order to reveal not only test sensitivity but the extent to which the test itself is "ecologically valid."

Finally, it is also important to utilize various behavioral measures in the assessment of intelligibility. Recently, some researchers (Manous, Pisoni, Dedina, and Nusbaum, 1985; Pisoni and Dedina, 1986; Schmidt-Nielsen and Kallman, 1987) have measured response latency in sentence-verification tests, reasoning that such tests are especially sensitive indicators of the cognitive "cost" involved in processing acoustically degraded stimuli. Whether this assertion is true may depend, in part, upon the type of nontimed intelligibility tasks with which timed tasks of sentence verification are compared. But the search for sensitive test procedures and materials is laudable and should continue. As the quality of synthetic speech improves, the need for such tests will grow.

In conclusion, the present study was designed to address several major questions about the intelligibility of natural and vocoded speech presented to native and non-native listeners of English. It is hoped that, in the process of addressing these questions, the experimenters have not only provided some important answers, but have also revealed the need for continued work directed at investigating speech processing among native and non-native listeners.

5. SUMMARY

An experiment was conducted in order to explore three main questions concerning the intelligibility of natural and LPC-vocoded linguistic stimuli presented to native English listeners and German-English bilingual (non-native) listeners. These questions were as follows: (1) To what extent is intelligibility affected when vocoded versus natural speech is presented to native and non-native listeners? (2) What is the magnitude of between-test differences when various types of intelligibility tests are utilized with these two groups? (3) Are specific patterns of response similar for native and non-native listeners? In order to answer these questions, the experimenters presented a natural-speech version of the DRT, MST, and SAST to ten native and ten non-native listeners, and a vocoded-speech version of the same tests to another group of ten native and ten non-native listeners.

Results revealed that the non-native listeners performed significantly worse than the native listeners in the vocoded condition on the DRT and in the natural and vocoded conditions on the MST and SAST. Moreover, the effects of various conditions upon response accuracy were nonadditive. In addition, the non-native listeners appeared to utilize different processing strategies from the natives, and they experienced greater recall difficulty. Word frequency affected response accuracy for both subject groups, though more so for the non-native than for the native listeners. Finally, unlike the performance of the native listeners, the performance of the non-native listeners on the natural version of the SAST and on the vocoded version of the MST and SAST worsened over the course of the test, possibly revealing the effect of test fatigue.

These results thus revealed significant differences not only in the intelligibility of natural and vocoded speech and in the overall performance of native and non-native listeners, but in the types of response strategies and recall and learning effects evinced by native and non-native listeners. Perhaps most important, the present study suggests that even moderate amounts of "perceptual loading" in a speech signal can induce serious decrements in intelligibility among non-native listeners, even when those listeners are highly fluent in the language presented.

APPENDIX A
Diagnostic Rhyme Test Words

gob	coot	test	pan
taunt	pond	vault	jock
boot	moan	news	dote
cheat	bill	vee	thick
gab	jest	sank	care
tot	thought	wad	bong
boast	poop	show	you
nill	reap	rip	neath
zed	fast	tense	gaff
daw	dock	moss	mom
choose	doze	foo	though
cheap	thing	thee	jilt
bank	net	Thad	pent
dot	taught	hop	yawl
rose	nude	node	roose
tint	bean	gin	feel
deck	bad	mend	nab
thong	vox	chaw	bon
coo	go	juice	sole
reed	bid	peak	thin
shag	wren	bat	keg
rob	naught	not	raw
foal	Sue	goat	dune
nip	need	bit	beat
fence	than	den	Chad
saw	chop	gauze	got
pool	fore	noon	dole
yield	fit	tea	gill
Nat	nest	rap	red

Bob	toot	pest	fan
daunt	bond	fault	chock
moot	bone	dues	note
sheet	vill	bee	tick
jab	guest	thank	chair
pot	fought	rod	dong
ghost	coop	so	rue
rill	neap	nip	wreath
said	vast	dense	calf
naw	knock	boss	bomb
shoes	those	pooh	dough
keep	sing	zee	gilt
dank	met	fad	tent
got	caught	fop	wall
nose	rude	rode	noose
dint	peen	chin	veal
neck	mad	bend	dab
tong	box	shaw	von
chew	Joe	goose	thole
weed	did	teak	fin
sag	yen	gat	peg
knob	wrought	rot	naw
vole	zoo	coat	tune
dip	deed	mitt	meat
pence	Dan	then	shad
thaw	cop	jaws	jot
tool	Thor	moon	bowl
wield	hit	key	dill
rat	rest	nap	Ned

APPENDIX B
Meaningful Sentences

1. The giddy children mimicked a silly theme.
2. Shy Laura shares the Chinese food.
3. Gutsy Hank tames a vicious tiger.
4. A shifty burglar hides the magic zither.
5. The happy girl feeds the thirty turtles.
6. A nervous mother feels the sharp knife.
7. A fearless ranger hurt a hungry bear.
8. Cautious Vickie tastes the salty chili.
9. A caring doctor tends the needy patients.
10. The lush zoo contains a tame chimp.
11. Pious John pardons the sinful people.
12. The furry puppy chews a mink jacket.
13. The fair judge changed a harsh verdict.
14. Zestful Mary digs a deep hole.
15. The tidy husband dusts the dirty house.
16. A noble family values a rare diamond.
17. The gallant leader vetoed the last bill.
18. The jilted chap left a festive show.
19. A rich captain hates a junky sailboat.
20. The big farmer lifts a bulky load.
21. Joyful Sally rides a copper bike.
22. Peppy Tom purchased a lively kitten.
23. The thoughtless cook damaged the lovely dessert.
24. Dull Carl bores a tired nephew.

25. The careless nurse zonked the sore thumb.
26. Shaky Libby jammed a rusty zipper.
27. A veiled gypsy jingles the zircon rings.
28. Bold Gail kicked the mad dog.
29. Chubby Nancy made the chocolate pudding.
30. A lonely child cherished a gentle pet.
31. A native savage captured the holy man.
32. A jeering bum gobbles a cheese sandwich.
33. A tender guy wrapped the burned finger.
34. The fast guard seizes a harmful thug.
35. A dauntless juggler gathers the fallen marbles.
36. A daring pilot zooms a mighty jet.
37. Valiant Zelda saved a sinking ship.
38. A sincere teacher nurtures a vital theory.
39. The busy neighbor mowed the massive lawn.
40. The royal court regrets the shameful news.
41. A kindly dentist numbs a painful tooth.
42. The vexed bull bucks a rowdy cowboy.
43. Patient Vance fattened the hairy goats.
44. A tawdry maid vacuumed the shag rug.
45. The polite lady shunned a naughty thought.
46. The jumpy zealot thumped a mean vandal.
47. A gifted seamstress sews a zippered shawl.
48. A nimble rancher sheared a thousand sheep.
49. The vocal valet nags the cheerless duke.
50. Thankful Hans loves the cheery music.

51. Godly Zeus zaps the raging thunder.

52. Dismayed Vincent jinxed the baseball game.

53. A Zulu fellow pulls a tan horse.

54. The cagey fighter got the largest reward.

55. A thermal tide bathes the zonal chill.

56. The thankless chef thins the pea soup.

57. The gainly shepherd thatches a zigzag roof.

APPENDIX C
Semantically Anomalous Sentences

1. A painted shoulder thawed the misty sill.
2. The bitter seed vexes a valid dinner.
3. The tacky runner judged a short fact.
4. Dingy Doug chips the poor jewel.
5. A golden corner varies the thoughtful keeper.
6. A cotton zebra thickened the chief tickle.
7. The simple rocket picks a new female.
8. A zesty joke gets the nice feather.
9. The shiny shore gives a heavy father.
10. Checkered Sharon gained the chilly hope.
11. Recent Gary sets a messy shower.
12. Fake Chuck finished the hopeful golfer.
13. The vague job savors a jolly garden.
14. A thin jailer checked a meager soap.
15. Moody Tim holds the sane zero.
16. A newer deed shines a safe sinner.
17. A luscious devil helps the good raid.
18. The jealous duster lifted a gaudy cap.
19. The helpful knitter makes a gabby lip.
20. A paper nature seeks the cool master.
21. The bossy vapor shakes a careful victor.
22. Top Jane zapped the tense tot.
23. A dark nail zones the round reason.
24. The kind ladder shoots a dim bed.

25. The gilded nest zipped the dusty tank.

26. The zingy thing liked a late toddler.

27. The soft bargain mixes a thick needle.

28. A shoddy lobby mopped the dense hip.

29. Modern Leslie healed a cheap hat.

30. The charming deck robbed the hot jelly.

31. A jaunty fork raised a vacant cow.

32. The funny heaven reads the shallow pepper.

33. Ready Holly doubts the shabby van.

34. Novel Cathy dipped the loud hopper.

35. A vain foam denies a zippy lime.

36. The third pattern teases a zany tailor.

37. High Mick thanked a zealous chin.

38. Healthy Ned tears the solid rat.

39. Lean Rex takes the pale chowder.

40. A lewd pill leads a pink zing.

41. The bizarre pot needed the best zombie.

42. A partial baker knocked the boring shell.

43. Tipsy Peter keeps the better chopper.

44. The damp vase catches a tiny zeal.

45. A kingly thinker bites a nasty lock.

46. A gorgeous villain chopped the rotten thimble.

47. The southern gift beats the tall thighs.

48. Sure Susan bought a famous thirst.

49. A jagged sailor paid a ripe card.

50. A cheerful thistle pours the fat bean.

51. The zinc mitt carries a lazy basket.

52. A feisty chain fights the fertile money.

53. Vast Bob jabbed a junior pack.

54. The thirsty vine finds a giant shop.

55. The moral gold vacates a costly gate.

56. A normal cheater joined the thorough mess.

57. Rapid Zach nabs a vulgar mirror.

REFERENCES

Boothroyd, A. and Nittrouer, S. 1988. Mathematical treatment of context effects in phoneme and word recognition. The Journal of the Acoustical Society of America, 84, 101-114.

Clark, J.E., Dermody, P., and Palethorpe, S. 1985. Cue enhancement by stimulus repetition: Natural and synthetic speech comparisons. The Journal of the Acoustical Society of America, 78, 458-462.

Cziko, G. 1980. Language competence and reading strategies: A comparison of first- and second-language oral reading strategies. Language Learning, 30, 101-114.

Gold, B. and Tierney, J. 1983. Vocoder analysis based on properties of the human auditory system. MIT Lincoln Laboratory Technical Report, No. 670, 22 December 1983.

Greene, B.G., Pisoni, D.B., and Gradman, H.L. 1985. Perception of synthetic speech by nonnative speakers of English. Research on Speech Perception, Indiana University, Progress Report No. 11, 419-428.

Gronlund, N.E. 1985. Measurement and Evaluation in Teaching. New York: MacMillan Pub. C. 5th Ed.

Grosjean, F. 1980. Spoken word recognition processes and the gating paradigm. Perception & Psychophysics, 28, 267-283.

Hoover, J., Reichle, J., van Tasell, D., and Cole, D. 1987. The intelligibility of synthesized speech: ECHO II versus VOTRAX. Journal of Speech and Hearing, 30, 425-431.

Kalikow, D.N., Stevens, K.N., and Elliot, L.L. 1977. Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. The Journal of the Acoustical Society of America, 61, 1337-1351.

LeBlanc, R. and Painchaud, G. 1985. Self-assessment as a second language placement instrument. TESOL Quarterly, 19, 673-687.

Luce, P.A., Feustel, T.C., and Pisoni, D.B. 1983. Capacity demands in short-term memory for synthetic and natural speech. Human Factors, 25, 17-32.

Mack, M. 1988. Sentence processing by non-native speakers of English: Evidence from the perception of natural and computer-generated anomalous L2 sentences. Journal of Neurolinguistics, 3, 293-316.

Mack, M. and Gold, B. 1985. The intelligibility of non-vocoded and vocoded semantically anomalous sentences. MIT Lincoln Laboratory Technical Report, No. 703, 26 July 1985.

Mack, M. and Tierney, J. 1987. The intelligibility of natural and vocoded semantically anomalous sentences: A comparative analysis of English monolinguals and German-English bilinguals. MIT Lincoln Laboratory Technical Report, No. 792, 10 December 1987.

Marslen-Wilson, W.D. and Welsh, A. 1978. Processing interactions and lexical access during word recognition in continuous speech. Cognitive Psychology, 10, 29-63.

McLeod, B. and McLaughlin, B. 1986. Restructuring or automaticity? Reading in a second language. Language Learning, 36, 109-123.

Miller, G.A., Heise, G.A., and Lichten, W. 1951. The intelligibility of speech as a function of the context of the test materials. Journal of Experimental Psychology, 41, 329-335.

Mullenix, J.W., Pisoni, D.B., and Martin, C.S. 1988. Some effects of talker variability on spoken word recognition. The Journal of the Acoustical Society of America, 85, 365-378.

Nye, P.W. and Gaitenby, J.H. 1973. Consonant intelligibility in synthetic speech and in a natural speech control. Haskins Laboratories, Status Report on Speech Research, SR-33.

Nye, P.W. and Gaitenby, J.H. 1974. The intelligibility of synthetic monosyllabic words in short, syntactically normal sentences. Haskins Laboratories, Status Report on Speech Research, SR-37/38.

Ozawa, K. and Logan, J.S. 1989. Perceptual evaluation of two speech coding methods by native and non-native speakers of English. Computer Speech and Language, 3, 53-59.

Pisoni, D. B. and Hunnicutt, S. 1980. Perceptual evaluation of MITalk: The MIT unrestricted text-to-speech system. IEEE International Conference Record on Acoustics, Speech and Signal Processing, 572-575.

Pisoni, D.B. and Dedina, M.J. 1986. Comprehension of digitally encoded natural speech using a sentence verification task (SVT): A first report. Research on Speech Perception, Indiana University, Progress Report No. 12, 3-18.

Pisoni, D.B. and Koen, E. 1982. Some comparisons of intelligibility of synthetic and natural speech at different speech-to-noise ratios. The Journal of the Acoustical Society of America, Suppl. 1, 71, S94.

Pisoni, D.B., Manous, L.M., and Dedina, M.J. 1987. Comprehension of natural and synthetic speech: II. Effects of predictability on the verification of sentences controlled for intelligibility. Computer Speech and Language, 2, 303-320.

Pollack, I. 1959. Message repetition and message reception. The Journal of the Acoustical Society of America, 31, 1509-1515.

Salasoo, A. and Pisoni, D.B. 1985. Interaction of knowledge sources in spoken word identification. Journal of Memory and Language, 24, 210-231.

Savin, H.B. 1963. Word-frequency effect and errors in the perception of speech. The Journal of the Acoustical Society of America, 35, 200-206.

Schmidt-Nielsen, A. and Kallman, H.J. 1987. Evaluating the performance of the LPC 2.4 kbps processor with bit errors using a sentence verification task. Naval Research Laboratory, NRL Report 9089.

Tremain, T.E. 1982. The government standard linear predictive coding algorithm: LPC-10. Speech Technology, 1, 40-43.

Voiers, W.D. 1977. Diagnostic evaluation of speech intelligibility. In M.E. Hawley (Ed.). Speech intelligibility and speaker recognition. Stroudsburg, PA: Dowden, Hutchinson and Ross. pp. 374-387.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	5 July 1990	Technical Report

4. TITLE AND SUBTITLE

The Intelligibility of Natural and LPC-Vocoded Words and Sentences Presented to Native and Non-Native Speakers of English

5. FUNDING NUMBERS

C — F19628-90-C-0002

6. AUTHOR(S)

M. Mack, J. Tierney, and M.E.T. Boyle

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Lincoln Laboratory, MIT
P.O. Box 73
Lexington, MA 02173-9108

8. PERFORMING ORGANIZATION REPORT NUMBER

TR-869

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

HQ AF Systems Command AFSC/XTKT
Andrews AFB
Washington, DC 20334-5000

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

ESD-TR-90-040

11. SUPPLEMENTARY NOTES

None

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

The experiment reported in the present study was designed to compare the intelligibility of natural and LPC-vocoded linguistic stimuli presented to native and non-native speakers (listeners) of English. Subjects were 20 native speakers of English and 20 native speakers of German who were fluent in English. Three types of stimuli — the Diagnostic Rhyme Test, the Meaningful Sentences Test, and the Semantically Anomalous Sentences Test — were presented in both natural and vocoded conditions.

Results revealed the following: (1) The non-native listeners performed significantly worse than the native listeners in the vocoded condition on the DRT and in the natural and vocoded conditions on the two sentence tests; (2) the effects of listening condition and test type upon response accuracy were nonadditive; (3) the non-native listeners appeared to utilize *processing strategies* unlike those of the native listeners; (4) the non-native listeners experienced greater recall difficulty than the natives; (5) word frequency affected response accuracy for both subject groups, though somewhat more so for the non-native than for the native listeners; and (6) unlike the native listeners the non-native listeners appeared to exhibit fatigue effects in response to vocoded speech.

These findings provide insight into the role of listening condition and test type in tasks of speech intelligibility, and they reveal differences in the types of response strategies and perceptual learning evinced by native and non-native listeners. In addition, the present study reveals that even moderate amounts of "perceptual loading" can result in serious intelligibility problems for non-native listeners — even when such individuals are quite fluent in the language presented.

14. SUBJECT TERMS

intelligibility	vocoded speech	semantically anomalous sentences
vocoder (vocoded)	Diagnostic Rhyme Test (DRT)	native listeners (speakers)
computer-generated speech	meaningful sentences	non-native listeners (speakers)

15. NUMBER OF PAGES

66

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	